International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 5, October 2018, pp. 4042 – 4046
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i5.pp4042-4046

Multi-modal Asian Conversation Mobile Video Dataset for Recognition Task

Dewi Suryani, Valentino Ekaputra, and Andry Chowanda
Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

Article Info

Article history:
Received December 6, 2017
Revised July 15, 2018
Accepted August 12, 2018

Keyword:
Multi-modal dataset
Asian conversation dataset
Mobile video
Recognition task
Facial features expression
Emotion recognition

ABSTRACT

Images, audio, and videos have been used by researchers for a long time to develop several tasks regarding human facial recognition and emotion detection. Most of the available datasets usually focus on either static expression, a short video of changing emotion from neutral to peak emotion, or difference in sounds to detect the current emotion of a person. Moreover, the common datasets were collected and processed in the United States (US) or Europe, and only several datasets were originated from Asia. In this paper, we present our effort to create a unique dataset that can fill in the gap by currently available datasets. At the time of writing, our datasets contain 10 full HD (1920×1080) video clips with annotated JSON file, which is in total 100 minutes of duration and the total size of 13 GB. We believe this dataset will be useful as a training and benchmark data for a variety of research topics regarding human facial and emotion recognition.

Corresponding Author:
Dewi Suryani, Valentino Ekaputra, Andry Chowanda
Computer Science Department, School of Computer Science, Bina Nusantara University
Jl. K. H. Syahdan No. 9, Palmerah, Jakarta, Indonesia 11480
(+6221) 534 5830, ext. 2188
{dsuryani, vekaputra, achowanda}@binus.edu

1. INTRODUCTION

Nowadays, a mobile device such as a smartphone is equipped with a high quality camera and a good processor which enabled people to record a high definition (HD) video. There is no need to buy an expensive digital or DSLR camera for the people who want to record a video with a good quality result. Based on the observation performed by Ofcom [1] in 2016, in the USA, people spent approximately 87 hours on average a month to browse on a smartphone compared to 34 hours on laptop or desktop. This signifies that people tend to use a smartphone for most of their activities. It means that the feature on their smartphone is more than enough for their daily needs, and this includes the video recording. Statista [2] also claimed that 71% of 44,761 respondents use their smartphone to take photos/videos, which is the second highest activity in the smartphone after accessing the internet. This indicates the camera in their smartphones is already satisfied for taking images and recording video compared to a few years before.

Despite the increasing amount of time people spend on their smartphone, there is no publicly available dataset regarding Asian facial feature that is captured using mobile smartphone camera. Generally, the dataset only exists still as images and most of the video dataset only covers the western person facial features. As literature suggested, although facial expression recognition is universal to all races of humans, emotions perception from facial expressions cues are quite different from one culture to others. Moreover, most of the datasets focus only on facial features of the video and there is no such thing as facial features when a person is talking with the others in the wild. Hence, by using smartphone, we could capture a natural conversation of two interlocutors in the wild. Such datasets serve as datasets for computer to learn emotions recognition, facial expression recognition features and classification.

In this work, we are aiming to address this gap by presenting a mobile video dataset that contains

videos of Asian people having a natural conversation with each other. Our dataset is obtained by recording a natural conversation between 2 people inside a controlled room with adequate lighting and the mobile camera is in a steady position. In order to collect the variety of facial features, we provide several topics to be chosen by the interlocutors. The topics are mostly general topics such as foods, lecturers, etc.

The rest of this paper is organized as follows: In Section 2., we list the existing different publicly available dataset and explain its difference with our dataset. Our method of collecting data and the characteristic of the data will be explained in Section 3. Potential applications of the dataset are described in Section 4. And lastly, the conclusion will be provided in Section 5.

2. RELATED WORKS

Currently, several datasets have been created for many kinds of recognition tasks, especially facial expression analysis and recognition [3, 4, 5]. However, only few datasets that contain Asian people and recorded by using a smartphone. The extended Cohn-Kanade database, CK+ [6] composed of 593 recordings of posed and non-posed sequences. It is recorded under controlled conditions of light and head motion, and range between 9-60 frames per sequence. Each sequence represents a single changing facial expression that starts with a neutral expression and ends with a peak expression. The transitions between expressions are not included. Moreover, there is an NRC-IIT database [7] which contains pairs of short low-resolution mpeg1-encoded video clips. Each video clip is showing a face of a user who sits in front of the monitor that exhibiting a wide range of facial expressions and orientations as captured by a USB webcam mounted on the computer. Every video clip is about 15 seconds long, has a capture rate of 20 fps and is compressed with the AVI Intel codec 481 Kbps bit-rate.

In the other hands, there is the Cohn-Kanade DFAT-504 dataset [8] that consists of 100 university students ranging in age from 18 to 30 years. 65% were female, 15% were African-American, and 3% were Asian or Latino. Students were instructed by an experimenter to perform a series of 23 facial expressions. Students began and ended each display with a neutral face. Image sequences from neutral to target display were digitized into 640 by 480 pixel arrays with 8-bit precision for grayscale values. Similar to the others, the MMI database [9] contains a large collection of FACS coded facial videos. However, it consists of 1395 manually AU coded video sequences with the majority of the video is posed and recorded in laboratory settings.

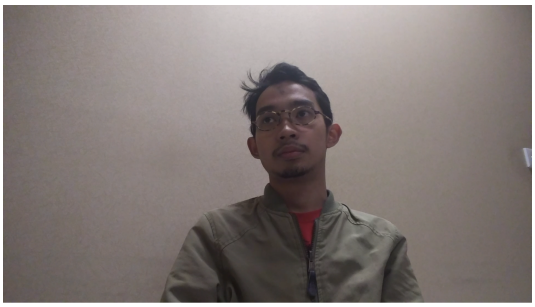
In the dataset mentioned above, most of them only focus on the image part of the video without recording the audio and makes the dataset quite unnatural for several expression. The dataset above mostly used camera or webcam to take the video. Also the duration of each data considered to be too short for applications in real life condition which combined different aspects and the context of the topic with the facial expression in the video. In summary, our dataset is different in following points: (i) recorded using the camera in a mobile device; (ii) long duration of natural conversation video; (iii) includes full HD videos; and (iv) includes audio for matching expression with context.

3. PROPOSED METHOD

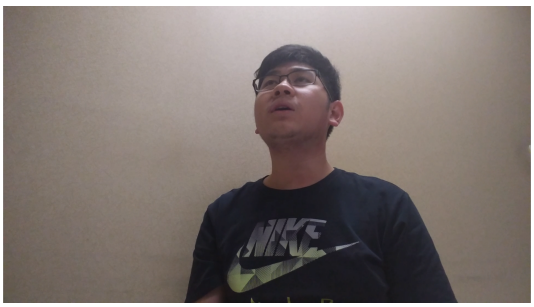
In this paper, we proposed a new mobile video dataset, which can be used as a benchmark data for several recognition tasks as well as serve as a dataset for machine learning tasks. Here, we start by explaining how we collect the dataset for this research. The aim of this research is providing a publicly available dataset for Asian (specifically Indonesian) facial features, expressions, and conversations. Thus, we recruited twenty volunteers, whose mainly are Indonesian students (age between 19 - 21) to participate in this research. The participants are given a list of possible topics to be discussed during the recording in 10 minutes. In one session of recording, there were two interlocutors sitting facing each other across the table. The participants then start the conversation in with the other interlocutor, when the researcher give signal to them to start the conversation.

To record the conversation, the researchers set up 2 smart phone with identical camera specification. The smart phone used in this research were two Xiaomi Mi 4i with 13MP, f/2.0 camera and the video was recorded in full a HD setting with resolution of 1920x1080 pixels and 30 fps. The smart phone were placed in a steady position in front of each interlocutor. The recorded video is depicted in the following Figure 1 where Figure 1a shows the first interlocutor who involves in this conversation with a specific topic selected, with the other interlocutor is illustrated in Figure 1b.

During the conversation, the volunteers are encouraged to behave as if it is a normal and natural conversation in order to get the most natural dataset possible. After 10 minutes, they will be reminded to stop the conversation and the video will be saved as an mp4 file with the MPEG-4 format in the device before it is ex-



(a) First Interlucutor



(b) Second Interlucutor

Figure 1. Example of recorded video in a conversation between two interlocutors


```
{
  "frames":{
    "1":[{
      "x1":293,
      "y1":56,
      "x2":469,
      "y2":243,
      "id":0,
      "width":746,
      "height":419,
      "type":"Rectangle",
      "tags":["neutral"],
      "name":1,
      "blockSuggest":true
    }],
    "2":[{
      "x1":295,
      "y1":72,
      "x2":471,
      "y2":259,
      "id":1,
      "width":746,
      "height":419,
      "type":"Rectangle",
      "tags":["neutral"],
      "name":1,
      "suggestedBy":{"frameId":1, "regionId":0},
      "blockSuggest":true
    }],
```


Figure 2. Part of annotated JSON example

ported to a computer. Each file will be named in the format as follows: "CONVERSATIONID_CAMERAID", where CONVERSATIONID is the identifier for the session and CAMERAID is the identifier for the device used. After the data is completely recorded, we start annotating the video for three different facial expressions, i.e., sad, neutral, and happy. Using visual object tagging tool (VOTT) [10] provided by Microsoft, we can get the annotation data in JSON format as shown in Figure 2.Furthermore, Figure 3 also describes the video example while annotating the data.

4. POTENTIAL APPLICATIONS

There are several potential applications that can take the advantages of the availability of this dataset, such as facial recognition and emotion detection.

Facial Recognition, our dataset provides another data in order to increase the accuracy of facial recognition task. The reason is that most research in this field is using CK+ dataset as done by Bartlett et al. [11], Cohen et al. [12], and Cohn et al. [13]. Unfortunately, the current used datasets mostly contain people from the western region, which can highly affect the Asian facial recognition.

Emotion Detection, most of other research regarding emotion detection only used specific datasets like visual only or audio only [14]. Our dataset provides multi-modal information, i.e., images and audio that linked together in this case. Moreover, the common research used is posed expression datasets that are not based on authentic emotions [15].

Virtual Humans or Intelligent VIrtual Agents, with the dataset, we can learn a natural conversation between two interlocutors and implement them into a virtual human [16, 17]. The dataset provides several features to be learn: emotion recognition from audio (i.e. voice) and audio (e.g. facial expressions), natural

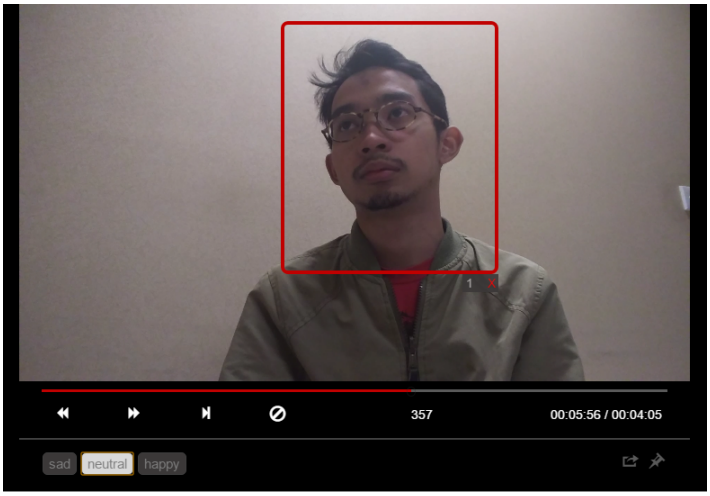


Figure 3. Example of the video while annotation process. The red square denotes the facial area of annotation.

language processing, and conversation.

Psychology or Social Science Study, with the dataset, researchers from psychology or social study also could analyze and observe human behavior during the interaction. An ethnography study also can be applied to analyze or to observe human behavior (specifically Indonesian people) through the video.

5. CONCLUSION

This paper presents conversation video dataset, containing videos of real conversation performed by a pair of volunteers recorded using mobile device camera along with JSON data of its annotation. In future, we intend to collect more data for the datasets by asking for more diverse volunteers based on age, gender, and occupation. We believe that this dataset will be useful for several applications which required training using images, audio, or videos from our datasets. We want also to record the video using different conditions of lighting in order to observe the influence of the lighting.

REFERENCES

[1] Ofcom. (2016, Dec.) The communications market report: International. [Online]. Available: https://www.ofcom.org.uk/research-and-data/multi-sector-research/cmr/cmr16/international

[2] Statista. (2016, Sep.) Weekly smartphone activities among adult users in the united states as of august 2016. [Online]. Available: https://www.statista.com/statistics/187128/leading-us-smartphone-activities/

[3] H. K. Palo and M. N. Mohanty, "Classification of emotional speech of children using probabilistic neural network," *International Journal of Electrical and Computer Engineering*, vol. 5, no. 2, p. 311, 2015.

[4] F. E. Gunawan and K. Idananta, "Predicting the level of emotion by means of indonesian speech signal." *Telkomnika*, vol. 15, no. 2, 2017.

[5] F. Z. Salmam, A. Madani, and M. Kissi, "Emotion recognition from facial expression based on fiducial points detection and using neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 1, 2017.

[6] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[7] D. O. Gorodnichy, "Video-based framework for face recognition in video," in *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on*. IEEE, 2005, pp. 330–338.

[8] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 46–53.

[9] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research*

on Emotion and Affect, 2010, p. 65.
[10] Microsoft. (2017, Sep.) Vott: Visual object tagging tool. [Online]. Available: https://github.com/Microsoft/VoTT
[11] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction." in *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, vol. 5. IEEE, 2003, pp. 53–53.
[12] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and image understanding*, vol. 91, no. 1, pp. 160–187, 2003.
[13] J. F. Cohn, L. I. Reed, Z. Ambadar, J. Xiao, and T. Moriyama, "Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 610–616.
[14] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, vol. 1. IEEE, 1997, pp. 397–401.
[15] Y. Sun, N. Sebe, M. S. Lew, and T. Gevers, "Authentic emotion detection in real-time video," in *International Workshop on Computer Vision in Human-Computer Interaction*. Springer, 2004, pp. 94–104.
[16] A. Chowanda, P. Blanchfield, M. Flintham, and M. Valstar, "Erisa: Building emotionally realistic social game-agents companions," in *International Conference on Intelligent Virtual Agents*. Springer, 2014, pp. 134–143.
[17] ——, "Play smile game with erisa," in *IVA 2015, Fifteenth International Conference on Intelligent Virtual Agents*, 2015.

BIOGRAPHIES OF AUTHORS

Dewi Suryani is a Computer Science lecturer at Bina Nusantara University, Indonesia. She obtained Bachelor Degree in Computer Science from Bina Nusantara University in 2014. She just graduated from the Sirindhorn Thai-German Graduate School of Engineering (TGGS), King Mongkut's University of Technology North Bangkok (KMUTNB), Thailand, with Master of Engineering. Her researches are in fields of computer science, image processing, handwriting recognition, etc.

Valentino Ekaputra is a lecturer in Bina Nusantara University. He obtained Bachelor and Master Degree in Computer Science from Binus University in 2016. Currently, he still does not have fields he is focusing on researches.

Andry Chowanda is a Computer Science lecturer in Bina Nusantara University Indonesia and currently is a PhD Student in the University of Nottingham. His research is in agent architecture. His work mainly on how to model an agent that has capability to sense and perceive the environment and react based on the perceived data in addition to the ability of building a social relationship with the user overtime.